DOCUMENT RESUME

ED 228 298 TM 830 218

AUTHOR Smith, Richard M.
TITLE A Comparison of Rasch Person Analysis and Robust Estimators.
PUB DATE Apr 83
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Difficulty Level; *Error of Measurement; Estimation (Mathematics); Goodness of Fit; Guessing (Tests); *Latent Trait Theory; *Maximum Likelihood Statistics; Psychometrics; *Responses; Scores; *Testing Problems; Test Items; Writing Skills
IDENTIFIERS Jackknifing Technique; *Rasch Model; *Robustness

ABSTRACT

Measurement disturbances, such as guessing, startup, and plodding, often result in an examinee's ability being either over- or under-estimated by the maximum likelihood estimation employed in latent trait psychometric models. Several authors have suggested methods to lessen the impact of unexpected responses on the ability estimation process. This study uses simulated data to compare the standard maximum likelihood estimation with two forms of robust estimation, BIWEIGHT and AMJACK, and person analysis within the Rasch model. The results indicate that, although the two robust estimation procedures recover the generating parameters under certain conditions, in the presence of many forms of measurement disturbances they mask important information. Rasch person analysis has the advantage of not only providing a method of modifying the ability estimation procedure but also of providing a means of identifying the nature of the disturbance. (Author)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED228298

# A COMPARISON OF RASCH PERSON ANALYSIS AND ROBUST ESTIMATORS

Richard M. Smith

American College Testing Program

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R.M. Smith

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

April, 1983

TM 830 218

Paper presented at the annual meeting of the
American Educational Research Association
Montreal, April, 1983

## ABSTRACT

Measurement disturbances, e.g., guessing, startup, plodding, etc., often result in an examinee's ability being either over- or under-estimated by the maximum likelihood estimation employed in latent trait psychometric models. Several authors have suggested methods to lessen the impact of unexpected responses on the ability estimation process. This study uses simulated data to compare the standard maximum likelihood estimation with two forms of robust estimaion, BIWEIGHT and AMJACK, and Person Analysis within the Rasch model.

The results indicate that, although the two robust estimation procedures recover the generating parameters under certain conditions, in the presence of many forms of measurement disurbances they mask important information. Rasch person anaysis has the advantage of not only providing a method of modifying the ability estimation procedure but also of providing a means of identifying the nature of the disturbance.

# A COMPARISON OF
# RASCH PERSON ANALYSIS AND ROBUST ESTIMATORS

Aberrant response patterns, such as guessing or carelessness, have long been considered a source of potential error in the measurement process. One possible solution to this problem is the use of robust estimation (Wainer & Wright, 1980; Mislevy and Bock, 1982) to moderate the effect of the unexpected responses on the estimation of a person's ability. These procedures are based on the calculation of the probability of a correct response for any item/person interaction within the framework of a latent trait model. These procedures make no assumptions about the presence or absence of measurement disturbances, such as guessing, startup, carelessness, etc., nor do they attempt to classify the cause of the disturbances, but, rather, check each response pattern for unexpected responses and make an appropriate statistical correction. One advantage of these corrections is that they are applied at the individual, rather than group, level. Persons without unexpected responses will have their overall ability estimate unchanged by these procedures.

Another solution to the problem caused by the

presence of aberrant response patterns is to use all the information available about the items and the persons in a systematic analysis of each person's response pattern. This analysis seeks to identify the cause and magnitude of the measurement disturbance and is referred to as person analysis (Mead, 1976; Wright, 1977; Wright and Stone, 1979; Wright, 1980). This procedure employs a variety of comparisons between expected and observed scores to detect the measurement disturbances.

Once such disturbances have been detected, there are four possible steps that can be taken.

1) Rather than report a single overall ability, report several separate abilities, e.g., one for addition and one for multiplication, etc.
2) Modify the response string and re-estimate the person's ability, e.g., eliminate the 10 un-reached items at the end of a person's response pattern.
3) Decide that there is not enough information in the response pattern to report any score and retest the individual.
4) Decide that the error introduced into the estimation procedure by the measurement disturbance is small enough so that the impact to the person's total ability estimate is

marginal. This decision can be based on a comparision of the error introduced by the measurement disturbance with the standard error of measurement associated with each ability estimate.

## RASCH PERSON ANALYSIS

Using the Rasch Model there are three basic methods of summarizing the information contained in the residual (obserevd - expected score) (Wright, 1980; Smith 1982a). These are the total, between, and within fit statistics. The unweighted total fit statistic squares the residual, standardizes it, and sums it over all the item/person interactions. The unweighted total is

$$UT = \frac{1}{L} \sum_{i=1}^{L} \frac{(O_i - E_i)^2}{E_i(1 - E_i)} \quad , \qquad (1)$$

where O is the observed response, E is the expected response,and L is the number of items on the test.

The unweighted between fit statistic is a direct test of the "item-freed" measurement property of the Rasch Model. That is, if the data fit the model, the overall ability estimate should accurately predict the person's score on any subset of items. By comparing the person's predicted score ($\sum$ E) with the observed score ($\sum$ O) on any subset of items it is possible to test the fit

of the data to the model. The unweighted between fit statistic is

$$UB = \frac{1}{(J-1)} \sum_{j=1}^{J} \frac{\left( \sum_{i \in j}^{L_j} O_i - \sum_{i \in j}^{L_j} E_i \right)^2}{\sum_{i \in j}^{L_j} (E_i (1 - E_i))}, \qquad (2)$$

where J is the number of subsets and $L_j$ is the number of items in each subset.

Both the total and the between fit statistics are based on the overall ability estimate. To test the fit of a response to the subset of items, it is possible to estimate the person's ability on that subset alone and create an unweighted within-set fit statistic for each subset. This fit statistic is

$$UW = \frac{1}{L_j} \sum_{i=1}^{L_j} \frac{(O_i - E_{ij})^2}{(E_{ij}(1 - E_{ij}))}, \qquad (3)$$

where $E_{ij}$ is based on the subset ability estimate.

The evaluation of these summary statistics can be expedited by using a cube root transformation (Wright, 1980). See Smith (1982a) for corrections to eliminate the effects of using estimated abilities rather than known parameters in these statistics.

Each fit statistic plays a different role in the process of detecting measurement disturbances. The total fit statistic providing a single overall test of fit is

better at detecting guessing, sloppiness, and other types of random measurement disturbances.

The between fit statistic is better at detecting more systematic forms of measurement disturbances, e.g., startup, plodding, guessing to complete, and all disturbances resulting from item/person interactions. This test of fit requires the a priori development of non-overlapping subsets of items for each test, e.g., grouping the items on the basis of item difficulty, position on the test, item type, specific forms of bias, etc. It is possible to perform any number of between tests on each response pattern.

The within fit statistic is not very useful in detecting new instances of measurement disturbance. Its greatest utility is in helping to identify the possible causes of measurement disturbance detected by the other two fit statistics. It is possible to compute a within fit statistic for each subgroup used in the between fit analysis.

## ROBUST ESTIMATORS

Robust estimation is a statistical procedure that is designed to lessen the impact of unexpected responses on the estimation of the person's ability. Two such procedures will be investigated in this study, the

AMT-Robustified Jackknifed procedure proposed by Wainer and Wright (1980) and the Biweight procedure proposed by Mislevy and Bock (1982). These procedures will be compared with the person analysis approach, discussed earlier, using a variety of simulated data sets that represent various catagories of measurement disturbances.

The AMT-Robustified Jackknifed (AMTJACK) procedure is based on a combination of two previously developed techniques, the jackknife estimation procedure (Tukey, 1958) and the Sine M-estimator (AMT) (Andrews et al, 1972). AMTJACK uses the jackknife procedure to estimate a pseudo-value of the individual's ability for a test with a single item eliminated. The procedure is used to estimate L separate pseudo-values by systematically eliminating each item on the test. These pseudo-values are defined by

$$A_i^* = L(A_t) - (L-1)(A_{t-1}) , \qquad (4)$$

where A is the ability based on the entire test, A is the ability based on the test with one item omitted, and L is the number of items on the test.

AMTJACK then uses the Sine M-estimator to combine the L $A^*$ pseudo-values into a single ability estimate. This procedure has an influence function that is "nearly like that of the mean for observations in close but goes to zero at the extremes" (Wainer and Wright, 1980). The

procedure is applied separately to the positive and negative pseudo-values. The AMT estimates for the positive and negative pseudo-values are then weighted by the number of observations, summed, and divided by the total number of observations to produce the AMTJACK ability estimate.

The basic result of using this procedure on a response pattern that contains aberrant responses will be to regress the ability estimate toward the center (average item difficulty) of the test. The extent of the regression depends on the number of aberrant responses. The greater the number of aberrant responses, the less confidence can be placed in the overall ability estimate and the more the AMTJACK ability estimate moves toward the test center while the standard error of the ability estimate increases.

The BIWEIGHT estimator (Mislevy and Bock, 1982) is based on Tukey's Biweight, a robust estimator of location. The BIWEIGHT estimate of an individual's ability is found by solving the modified maximum-likelihood equation

$$\sum_{i=1}^{L} W_i (X_{vi} - P_{vi}) / S_i = 0 \quad , \qquad (5)$$

where $W = \begin{cases} (1-U_{vi}) & \text{for } |U_{vi}| < 1 \\ 0 & \text{otherwise} \end{cases}$

$$\text{and } U_{Yi} = \frac{d - \tilde{\theta}_Y}{C \cdot S_i} \quad .$$

In this equation $X_{Yi}$ is the response of person $Y$ to item $i$, $P_{Yi}$ is the probability of a correct response for person $Y$ to item $i$, and $\tilde{\theta}_Y$ is the estimated ability of person $Y$. In the Rasch Model $S_i$ (the discrimination index) is 1.0 and C is a constant.

Mislevy and Bock claim that this procedure has four advantages over the usual maximum-likelihood estimation procedure for estimating ability.

1) The nature and extent of the disturbances are not assumed to be the same for all individuals.
2) Each response is utilized in proportion to its apparent value.
3) The BIWEIGHT estimate equals the maximum-likelihood estimate when no disturbances are present in the response pattern.
4) Computation requires only a minor modification of the Newton-Raphson method of maximum-likelihood estimation.

The effect of this procedure is to weight the contribution of each response in a response pattern by a weight that is a function of the unexpectedness of the response. A correct response that is close to the predicted probability of a correct response will carry

more weight in the estimation procedure than a correct response to an item which the individual had a small probability of responding to correctly. The lessening of the weight for unexpected responses contrasts with the AMTJACK procedure which regresses the ability estimate to the center of the test in the presence of unexpected responses.

## ANALYSIS OF SIMULATED DATA

In order to compare the usefulness of the two robust estimators and person analysis in detecting and/or correcting for the presence of aberrant responses, 3 data sets were generated to simulate various forms of measurement disturbance. The simulated disturbances were guessing, startup, and content/person interactions. Each of these data sets contained 100 replications for a person with a given ability. In each case the simulated test was 50 items long. The item difficulties were taken from an existing test of writing skills. The average item difficulty was 0.0 logits and the range was from -2.46 to 1.84. Eight of the items had a difficulty less than -1.0, 20 fell between -1.0 and 0.0, 14 were between 0.0 and 1.0, and 8 had a difficulty greater than 1.0.

### Measurement Disturbance #1: Guessing

The generation of response patterns required the selection of an overall ability. The ability chosen for

all replications was -1.0 logits so as to increase the number of items more than two logits from the person's ability. For the 42 items with a difficulty less than 1.0 logits, the probability of a correct response was calculated and compared to a random number between 0 and 1. If the random number was equal to or less than the probability of a correct response, the response was set equal to 1 (correct); otherwise, it was set equal to 0 (incorrect). For the 8 items with a difficulty greater than 1.0, the probability of a correct response was set equal to .25. This is the probability of correctly answering if the person was actually guessing on these 8 items. If the person had been responding in accordance with his ability, the probability of a correct response for these 8 items would have been less than .12.

This situation closely resembles the "guessing when the answer is not known" condition, since random responses were entered for every item more than 2.0 logits above the person's ability. Thus, for each response string, 42 responses were generated to fit the Rasch model and 8 were random responses. This is the type of situation that both AMTJACK and BIWEIGHT were designed to handle. Two ability estimates for each of the 100 replications were obtained using each of the two robust estimators. In addition, a person analysis program, PARM

(Smith, 1981),was run to calculate the total, between, and within unweighted fit statistics. In this instance, the between-set analysis was based on item difficulty. The ranges of item difficulties in logits for the four subsets were: less than -1.0; 1.0 to 0.0; 0.0 to 1.0; and greater that 1.0. The ranges of these item groups will vary according to the distribution of item difficulties on the test under study and were chosen to give four approximately equivalent groups.

A summary of the ability estimates obtained using the three procedures is presented in Table 1.

TABLE 1
COMPARISON OF ABILITY ESTIMATES

| | BIWEIGHT | AMTJACK | Uncorrected MLE |
|---|---|---|---|
| Mean Ability | -0.95 | -0.81 | -0.86 |
| Std. Dev. | 0.37 | 0.32 | 0.35 |
| Avg. S.E. | ---- | 0.36 | 0.33 |

The introduction of unexpected correct responses arising from setting the probability of a correct response at 0.25 for the 8 most difficult items produced three patterns. The AMTJACK estimation procedure, upon encountering the unexpected responses, moved the mean ability closer to the origin of the ability scale and increased the standard error of the ability estimate.

This moved the mean ability estimate further away from the generating ability of -1.0. The BIWEIGHT estimation procedure placed less of an emphasis on the unexpected responses and moved the mean of the ability distribution closer to the generating value. The usual maximum-likelihood estimation procedure weighted all correct responses equally. Thus, the unexpected correct responses made those in the population appear somewhat more able than they were.

It should be noted that the error impact, the bias introduced into ability estimation by the unexpected responses, was rather small. The average bias introduced, .14 logits, was about 1/2 of a standard error of measurement for each person. In this situation, the BIWEIGHT estimation procedure recovered the generating ability parameter better than the other two procedures. The AMTJACK procedure recovered the ability parameter least well.

In reviewing the person fit statistics for the 100 replications in this sample, 37 cases had a total fit statistic greater than 2.0, and 26 cases had a between fit statistic greater that 2.0. (See Smith, 1982b, for a discussion of the choice of a critical value for these fit statistics.) In all, 39 cases had one or both of the two fit statistics greater than 2.0.

In person analysis, the usual procedure in cases where a measurement disturbance is detected is to remove the unexpected responses from the ability estimation procedure and reestimate the ability parameter based on the modified response string. In this situation the responses to the 8 difficult items were dropped and the ability parameter was reestimated on the basis of the 42 less difficult items. Thus, 37 cases with a fit statistic greater than 2.0 had the ability parameter reestimated. The 63 cases with a fit statistic less than 2.0 retained their original ability estimate. Basing the ability estimate on 42 rather than 50 items increased the standard error of measurement for the 37 cases from approximately .33 to .36. The average standard error for the other 63 cases remained unchanged.

With the 37 corrected ability estimates in the sample the mean ability of the 100 cases was recalculated. The results are presented in Table 2.

TABLE 2
COMPARISON OF ABILITY ESTIMATES

| | BIWEIGHT | AMTJACK | Corrected MLE |
|---|---|---|---|
| Mean Ability | -0.95 | -0.81 | -0.98 |
| Std. Dev. | 0.37 | 0.32 | 0.39 |
| Avg S.E. | ---- | 0.36 | 0.35 |

Thus, the maximum-likelihood estimation combined with person analysis will recover the generating parameters better than either of the two robust estimators in the case where guessing on difficult items is the measurement disturbance that has been simulated.

Measurement Disturbance #2: Startup

Startup is a term used to describe the situation that occurs when a person who has a high level of test anxiety performs poorly on the items at the beginning of the test, but, after overcoming the anxiety, goes on to perform at a uniformly higher level. Due to the anxiety, the person appears to be less able on the initial items and more able on the latter items. If the person's true ability is most accurately reflected by his performance on the latter items, the ability estimate based on the entire response pattern will underestimate his ability.

In order to simulate this condition, two different ability estimates were used to generate the 50 item response pattern. An ability of -2.0 logits was used to generate the responses to the first 8 items on the test. An ability of 1.0 logit was then used to generate the responses to the remaining 42 items. The mean difficulty of the first 8 items was -1.2 logits, indicating fairly easy items. The average probability of a correct response

on these 8 items for a person with an ability of -2.0 is .31. The between fit subsets for this analysis were based on the item's position on the test. The test was divided into four parts, item numbers 1-12, 13-25, 26-37, and 38-50.

Ability estimates were obtained for both of the robust estimators as well as the uncorrected maximum likelihood procedure. The results of these three are reported in Table 3.

TABLE 3
COMPARISON OF ABILITY ESTIMATES

| | BIWEIGHT | AMTJACK | Uncorrected MLE |
|---|---|---|---|
| Mean Ability | 0.67 | 0.48 | 0.52 |
| Std. Dev. | 0.37 | 0.29 | 0.31 |
| Avg. S.E. | ---- | 0.38 | 0.32 |

Once again, in the presence of unexpected responses,the AMTJACK procedure shifted the estimated ability away from the uncorrected maximum likelihood estimate towards zero and increased the standard error of the ability estimates. The BIWEIGHT procedure, by placing less emphasis on the unexpected responses, moved the ability estimate closer to the generating value of 1. The maximum-likelihood procedure gave a mean ability estimate of .52. The impact (mean generating ability minus mean

estimated ability) of the 8 startup items on the MLE ability estimates was considerably larger that of the 8 guessing items (.48 vs. .14 logits) in the discussion above.

This increase in the error impact due to the unexpected responses was also reflected in the person fit statistics calculated for the 100 replications in this sample. A total of 62 cases had a total fit statistic greater than 2.0, and a total of 57 cases had a between fit statistic greater than 2.0. In all, 74 of the cases had either a total or between fit statistic greater than 2.0.

In these cases the power of the fit statistics to detect the measurement disturbance was higher than in the previous simulations. This was in spite of the fact that the disturbance was limited to the first 8 items and that the first subset in the between fit analysis contained 4 additional items with responses that fit the model. The usual procedure in a case such as this, where the measurement disturbance is confined to a particular portion of the response string, is to disregard those items and base the new ability estimate on the items that were unaffected by the influence of the high level of test anxiety.

With the 74 corrected ability estimates included in the sample the mean and standard deviation of the ability estimates were recalculated. The results are presented in Table 4.

TABLE 4
COMPARISON OF ABILITY ESTIMATES

| | BIWEIGHT | AMTJACK | Corrected MLE |
|---|---|---|---|
| Mean Ability | 0.67 | 0.48 | 0.94 |
| Std. Dev. | 0.37 | 0.29 | 0.45 |
| Ave. S.E. | ---- | 0.38 | 0.39 |

The maximum-likelihood estimation procedure used in conjunction with person analysis recovered the generating parameters better than either of the robust estimators. It should be noted that the difference in the performance of the three methods in this case is considerably larger than in the previous case. This is primairly due to the fact that both robust estimators are designed to correct random responses. In the case of startup there is a systematic bias introduced into the estimation process. This type of systematic bias is more efficiently detected by person fit analysis.

Measurement Disturbance #3: Content/Preson Interaction

The final type of measurement disturbance that will be studied is the content/person interaction. The 50 item

test from which the item difficulties for these simulations were taken was composed of four different content areas. Items 1-12 were from one area, items 13-25 a second, items 26-37 a third, and items 38-50 a fourth. The mean item difficulties for these 4 subsets of items were -0.35, 0.49, -0.23, and 0.51 respectively. In order to simulate the condition where a person was unfamiliar with one of the content areas, due, perhaps, to a non-uniform curriculum, two different abilities were used to generate the responses. The ability used to generate the responses to the first three subsets (items 1-37) was 1.0 logit. The ability used to generate the responses to the fourth subset was -1.0 logits. For this subset the average probability of a correct response was .18, given the ability of -1 and the average item difficulty of .51. The between fit statistic subsets in this analysis corresponded to the item content groups mentioned above.

Ability estimates were obtained for both of the robust estimators as well as the uncorrected maximum likelihood procedure. These results are presented in Table 5.

TABLE 5
COMPARISON OF ABILITY ESTIMATES

| | BIWEIGHT | AMTJACK | Uncorrected MLE |
|---|---|---|---|
| Mean Ability | 0.48 | 0.44 | 0.50 |
| Std. Dev. | 0.34 | 0.32 | 0.34 |
| Avg. S.E. | ---- | 0.32 | 0.32 |

In this case the unexpected responses were not confined to the 8 most difficult items, as was the case in the guessing simulation, nor to the 8 easy items at the beginning of the test, as was the case in the startup simulation. The unexpected responses in this simulation were on a 12 item subset that ranged in difficulty from -1.09 up to 1.84 logits.

In the presence of this type of systemetic disturbance the AMTJACK procedure moved the mean ability estimate closer to zero.The magnitude of the shift was approximately equal to the shifts for the other disturbances, although there was no noticeable increase in the standard error of measurement in this simulation. For the first time the BIWEIGHT procedure moved the mean ability estimate away from the generating ability for the majority of the items (+1.0). The magnitude of the difference between the uncorrected MLE ability mean and that for the BIWEIGHT procedure is less than in the

previous two simulations, .02 vs. .09 and .15.

When these 100 replications were analyzed with the person analysis program, the total fit statistic for only one case was found to be larger than 2.0. This is somewhat surprising since none of the responses were generated to fit the average estimated ability and about 1/4 of the items were generated to fit an ability 1.5 logits lower than the average estimated ability. In fact, the mean total fit statistic for these 100 cases was -0.30 with a standard deviation of 0.99. However, a total of 52 cases had a between fit statistic greater than 2.0. The mean between fit statistic for these replications was 1.85 with a standard deviation of 1.06. A total of 64 cases had a between fit statistic larger than the less conservative critical value of 1.5. The data presented here support the belief that the total fit statistic is insensitive to various forms of systematic measurement disturbances.

With this type of measurement disturbance, it would not make any sense to report a single overall ability estimate or to attempt to correct the MLE ability estimates. Since these "individuals" have two different abilities on different content areas, the test can no longer be considered unidimensional for them. The best procedure to follow in cases like this is to report four

different ability estimates, one for each of the different content areas. It is clear that none of the cases can be accurately labeled as having an overall ability of approximately .46 logits as the two robust estimators would have us believe. The best that can be done in this case with person analysis is to keep from 52 to 64 of the 100 cases, depending on the critical value chosen for the between fit statistic, from being misrepresented by the overall ability estimate.

## CONCLUSIONS

These simulations support the hypothesis that maximum likelihood ability estimation, when combined with person analysis, will provide a more accurate representation of the examinee's ability than the two forms of robust estimation when some form of measurement disturbance is present. Since most measurement disturbances can be described as systematic in nature, it generally does not make sense to report a single overall ability estimate. The use of a robust estimator can actually mask important information about the person being measured, e.g., a particular content area in which the individual is deficient. In practical situations this deficiency could be corrected by a short remedial course, rather than by placing the student into a lower level course where the majority of the topics covered will be

below his level of ability. The use of robust estimators in estimating ability must, therefore, be approached with caution.

## REFERENCES


Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. & Tukey, J.W. Robust estimates of location. Princeton, N.J., Princeton University Press, 1972.

Mead, R.J. Analysis of fit of data to the Rasch model through analysis of residuals. Ph.D. Dissertation, University of Chicago, 1976.

Mislevy, R. & Bock, R.D. Biweight estimates of latent ability. Educational and Psychological Measurement, 1982, 42, 725-737.

Smith, R.M. Methods of assessing the fit of persons to the Rasch model. A paper presented at the annual meeting of the National Council on Measurement in Education, New York, 1982a.

Smith, R.M. Detecting measurement disturbances with the Rasch model. Ph.D. Dissertation, University of Chicago, 1982b.

Smith, R.M. PARM: Person fit analysis with the Rasch model. Basic computer program, University of Chicago, 1981.

Tukey, J.W. Bias and confidence in not quite large samples. Annals of Mathematical Statistics, 1958, 29, 614. (Abstract)

Wainer, H. & Wright, B.D. Robust estimation of ability in the Rasch model. Psychometrika, 1980, 45, 373-391.

Wright, B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B.D. Afterword in Probablistic models for some intelligence and attainment tests, by G. Rasch. Chicago: University of Chicago Press, 1980.

Wright, B.D. and Stone, M.H. Best test design. Chicago: MESA Press, 1979.